# Strategic Openendedness

Sandy Tanwisuth[*]

April 23rd, 2025

## 1 Motivation

What if the generalization observed in tightly constrained two-player zero-sum (2pzs) environments is not a side effect of the reward structure, but a **diagnostic signal** of robust strategic learning?

In these settings, adversarial pressure and game-theoretic feedback compel agents to develop representations that are both performant and structurally invariant. Only abstractions resilient to strategic interrogation tend to persist. By contrast, open-ended environments remove these pressures. They offer no fixed opponents, no clear objectives, and often no stable ground truth. This reveals the brittleness of hardcoded rewards and the limitations of supervised goal specification.

Rather than abandoning adversarial learning, we ask: what if we **recast** it?

Suppose generalization under pressure is not just a consequence, but a **design principle**.

> In open-ended, multi-agent environments where reward is unstable or absent, what should agents optimize to recover the kind of robustness seen in two-player zero-sum games?

We propose **influence** as a strategic analogue to reward. Not as control, but as the capacity of an agent to shape, and be shaped by, the evolving behavior of others. Influence becomes the operative currency of abstraction. Agents are no longer curiosity-driven explorers or empowerment maximizers. They are ***representation learners***, motivated to ***reduce ambiguity across the co-strategic space***.

Where conventional intrinsic rewards falter in social settings: often overfitting to novelty, noise, or uncoordinated exploration. We propose a structured alternative. Reward agents for becoming legible. Incentivize shifts in others' behavior that are compressible. Encourage the cultivation of influence that becomes easier to model over time.

This principle is tested in **MettaGrid**, a compositional environment where emergence is scaffolded by intentional action, spatial affordances, and inter-agent dependencies. In MettaGrid, agents are not evaluated solely on task completion. They must coordinate ***trajectories of influence***. Solving the puzzle is insufficient; agents must shape how their actions are interpreted and adapted to by others.

This reframes open-endedness. The core question is no longer:

> How do we maximize reward in an unbounded space?

But rather:

> What **abstractions** support strategically generative behavior when goals are undefined?

And just as in 2pzs, robustness here is not measured by task success alone. It is defined by an agent's ability to **withstand strategic scrutiny**.

[*] Electronic address: `sandy314@duck.com`; Corresponding author; we thank members of Softmax in particular Emmett Shear, Jack Heart, and David Bloomin for helpful conversations. Draft of a work-in-progress shared for transparency and attribution. Not affiliated with any organization. Authored independently. Please contact the first author at sandy314@duck.com for licensing, collaboration, or reuse.

## 2 Framing and Core Contributions

At the heart of this proposal is a conceptual inversion: strategic influence, defined as the legible and learnable modulation of others' responses, is treated as the primitive instead of reward. In this frame, agents are not goal-optimizers but interaction-shapers. Reward becomes diagnostic rather than definitive.

We refer to this perspective as ***Strategic Open-Endedness***. The key insight is that in open-ended multi-agent environments, the *stability* of influence over time holds greater significance than the specificity of goals. Influence, in our definition, is not about dominance but about compressibility: the degree to which an agent's behavior reduces uncertainty in future strategic responses. In this view, influence is regularity that can be learned.

To ground this framing, we draw on the **Unified Strategic Representation Learning Theory (USRLT)** [Tanwisuth 2025], a modular architecture structured around layered strategic abstraction.

**1. Influence as Intrinsic Reward**

We define intrinsic motivation not in terms of novelty-seeking or prediction error, but as the process of disambiguating co-player strategy spaces. Agents are rewarded for refining their compressions over Strategic Equivalence Classes (SECs), producing a learning signal $\Delta\text{SEC}(t)$ that tracks alignment through abstraction rather than through direct outcome matching.

**2. Modular Strategic Compression**

USRLT integrates four layers of strategic abstraction:

- *Intent Certainty (IC)*: Quantifies the legibility of a co-player's influence on the ego-agent's soft best responses.
- *Contrastive Strategic Coding (CSC)*: Encodes short-horizon incentives using InfoNCE-style losses over response similarity.
- *Successor Feature Equivalence (SFR-SER)*: Groups policies by their long-term impact on expected feature trajectories.
- *Model-Aware Strategic Refinement (MASR)*: Captures how co-players reshape the ego-agent's induced transition dynamics via model-based inference.

Together, these layers form a hierarchy from immediate incentives to broader environmental influence. Through compression across these layers, agents attain strategic clarity—not by seeking novelty, but by reducing ambiguity.

**3. A Curriculum over Influence**

We replace task-based curricula with a curriculum centered on influence. Agents are trained with co-players who are *ambiguously compressible*, encouraging them to maximize marginal gains in their representational abstraction. This strategy fosters generalization through contrastive exposure to challenging coordination.

**4. Implementation in MettaGrid**

We validate USRLT in the *MettaGrid* environment, where agents must infer intentions, affordances, and causal dependencies amid entangled and often deceptive influence. Our evaluation focuses on:

- $\Delta\text{SEC}(t)$ as a shaping signal for learning
- Layer-wise convergence across IC, CSC, SFR, and MASR
- Emergent specialization and self-modeling in response to sustained influence ambiguity

**5. Reframing Open-Endedness**

Where prior frameworks emphasize novelty, we emphasize strategic learnability. Novelty without compression leads to noise; open-endedness without abstraction causes drift. We offer a new axis for evaluating AGI: not only whether agents solve tasks, but whether they become *understandable*—strategically legible within a co-evolving ecosystem.

# 3 Technical Approach

This section outlines how we operationalize strategic open-endedness using the Unified Strategic Representation Learning Theory (USRLT) framework. Our goal is to build agents that do not maximize extrinsic reward per se, but instead maximize the clarity and generality of their influence on other agents through layered representation learning. To do this, we define an optimization objective over Strategic Equivalence Class (SEC) refinement, implemented modularly across four abstraction layers. These abstractions yield a learning signal, $\Delta\mathrm{SEC}(t)$, which serves as the intrinsic reward function.

## 3.1 Learning Objective: Maximizing Strategic Compressibility

Let $\pi_j$ be a co-player policy and $\pi_i$ the ego-agent's policy. We define the ego-agent's internal representation of $\pi_j$ at time $t$ as a compressed embedding $z_j^{(t)} \in \mathbb{R}^d$, produced by an encoder $f_\theta$ acting on trajectory data $\tau_j$.

Each layer $\mathcal{L}_k \in \{\mathrm{IC}, \mathrm{CSC}, \mathrm{SFR}, \mathrm{MASR}\}$ defines a distinct compression function and associated mutual information objective. The intrinsic reward at time $t$ is defined as:

$$\Delta\mathrm{SEC}_k(t) = H\left[P_{\mathrm{intent}}^{(t-1)}\right] - H\left[P_{\mathrm{intent}}^{(t)}\right],$$

where $P_{\mathrm{intent}}$ is the agent's belief distribution over SECs (e.g., policy clusters that induce equivalent responses), and $H$ denotes Shannon entropy.

The total intrinsic reward is a weighted sum across layers:

$$r_{\mathrm{intr}}(t) = \sum_k \lambda_k \cdot \Delta\mathrm{SEC}_k(t).$$

This reward can be used to train $\pi_i$ via standard reinforcement learning updates, replacing or augmenting any extrinsic reward.

## 3.2 Layer 1: Intent Certainty (IC)

**Objective:** Estimate the legibility of a co-policy by measuring mutual information between co-policy $\pi_j$ and ego-action $a_i$.

We estimate $I(a_i; \pi_j \mid s)$ using a variational approximation. Specifically, we train a classifier $q_\phi(\pi_j \mid a_i, s)$ using a contrastive InfoNCE-style loss:

$$\mathcal{L}_{\mathrm{IC}} = -\mathbb{E}\left[\log \frac{q_\phi(\pi_j \mid a_i, s)}{\sum_{\pi_j'} q_\phi(\pi_j' \mid a_i, s)}\right].$$

The IC score is the negative entropy of the classifier's output: $\Delta\mathrm{SEC}_{\mathrm{IC}} \propto -H[q_\phi]$.

## 3.3 Layer 2: Contrastive Strategic Coding (CSC)

**Objective:** Embed co-policies based on the soft best response distributions they induce.

Let $\pi_i(a \mid s, \pi_j)$ denote the ego-agent's response distribution. We learn an embedding $z_j = f_\theta(\pi_j)$ such that similar soft BRs yield similar embeddings. This is enforced using the contrastive loss:

$$\mathcal{L}_{\mathrm{CSC}} = -\mathbb{E}_{\pi_j^+, \pi_j^-} \log \frac{\exp(\mathrm{sim}(z_i, z_j^+)/\tau)}{\exp(\mathrm{sim}(z_i, z_j^+)/\tau) + \sum \exp(\mathrm{sim}(z_i, z_j^-)/\tau)},$$

where $\mathrm{sim}(\cdot, \cdot)$ is cosine similarity and $\tau$ is a temperature hyperparameter.

### 3.4 Layer 3: Successor Feature Equivalence (SFR-SER)

**Objective:** Group co-policies by their long-term impact on expected feature trajectories.

We define the successor feature for co-policy $\pi_j$ and ego-policy $\pi_i$ as:

$$\psi^{\pi_j}(s) = \mathbb{E}_{\pi_i}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t)\right], \quad s_0 \sim P(\cdot \mid \pi_j),$$

where $\phi(s)$ are learned features. Agents cluster co-policies based on similarity of their $\psi$ vectors, with $\Delta\mathrm{SEC}_{\mathrm{SFR}}$ defined via entropy reduction over these clusters.

### 3.5 Layer 4: Model-Aware Strategic Refinement (MASR)

**Objective:** Characterize how a co-policy alters the ego-agent's internal transition model.

Given a co-policy $\pi_j$, we train an induced model $\hat{T}^{\pi_j}$ and compare it to a baseline model $\hat{T}^{\emptyset}$:

$$\mathrm{MASR}_j = D_{\mathrm{KL}}\left(\hat{T}^{\pi_j} \parallel \hat{T}^{\emptyset}\right).$$

Embeddings are learned such that similar influence over model dynamics clusters together; entropy reduction over these representations yields $\Delta\mathrm{SEC}_{\mathrm{MASR}}$.

### 3.6 Strategic Refinement Loop and Curriculum

Training follows a cyclic loop:

1. Sample a co-player $\pi_j$ from a pool of policies with high abstraction ambiguity or high expected $\Delta$SEC.
2. Roll out joint trajectories $\tau = \{(s_t, a_i, a_j)\}$ and compute layer-wise intrinsic rewards.
3. Update $\pi_i$ using reinforcement learning with $r_{\mathrm{intr}}$.
4. Update abstraction layer encoders via supervised or contrastive losses.
5. Adjust co-player sampling to maximize expected refinement of strategic abstractions.

### 3.7 Evaluation Metrics

To assess learning, we track:

- **Layer convergence:** Agreement across IC, CSC, SFR, and MASR partitions.
- **Abstraction refinement:** Change in entropy over SEC distributions, $\Delta\mathrm{SEC}(t)$.
- **Disambiguation:** Reduction in uncertainty when exposed to novel co-policies.
- **Causal alignment:** Do interventions on $\pi_j$ shift $\pi_i$'s abstractions as expected?

## 4 Experimental Design in MettaGrid

To validate the Strategic Open-Endedness framework and the layered USRLT abstraction model, we conduct controlled evaluations in the MettaGrid environment. MettaGrid is a grid-based multi-agent world in which agents interact through co-habitation, shared resources, and spatial dynamics. The environment supports the emergence of embodied competencies—such as spatial reasoning, influence modeling, and coordination—through varied but structured layouts.

Rather than optimizing for task completion, our experiments are designed to test the agent's capacity to compress, refine, and generalize co-player abstractions over time. In particular, we evaluate whether agents trained under a $\Delta$SEC-based intrinsic reward develop meaningful strategic competencies—such as influence legibility, policy disambiguation, and model-aware adaptation—even in the absence of fixed external objectives.

## 4.1 Environment Design

Each MettaGrid layout consists of:

- **Agents:** 2–3 independent policies (including the learner) sharing space and access to dynamic tiles, object triggers, and causal gates.
- **Interaction Zones:** Spatial affordances (e.g., doors, shared levers, color-coded tiles) that require or afford indirect influence.
- **State Perturbation Events:** Interventions such as object swaps, delay channels, or fog-of-war perturbations that selectively reveal how agents respond under uncertainty.

We vary task dynamics by controlling the opacity and ambiguity of co-player policies:

- **Fixed-policy co-players:** Hand-authored behaviors with known (but undisclosed) strategic patterns (e.g., greedy collector, delay maximizer).
- **Meta-learned co-players:** Policies sampled from a meta-trained family that generalizes across reward functions, thus appearing ambiguous unless disambiguated via extended interaction.

## 4.2 Experimental Conditions

We consider the following structured experimental blocks, aligned with embodied objectives:

- **Influence Clarification (IC):** Agents are paired with ambiguous co-policies. The goal is not to coordinate, but to reduce posterior uncertainty over co-player intention via policy compression.
- **Strategic Disambiguation (CSC):** Agents must choose between multiple affordances, where the optimal choice depends on latent co-player types (e.g., cooperative vs. deceptive). The objective is to cluster these latent strategies via induced best responses.
- **Successor Feature Evaluation (SFR):** Agents are assessed on their ability to construct temporally predictive embeddings of co-player behavior, grounded in long-run shared feature visitation.
- **Model Refinement (MASR):** We alter the environment's transition dynamics based on the co-player's behavior (e.g., movement over trigger tiles alters lava spread or door logic). The ego-agent must detect and refine its model of the world's causal graph conditional on co-player strategy.

Each episode includes a mixture of shared-space coordination opportunities and solo episodes for evaluation of strategy-internal consistency.

## 4.3 Metrics and Evaluation

We assess performance across the following axes:

- $\Delta$SEC **per timestep:** The reduction in entropy over Strategic Equivalence Classes at each layer (IC, CSC, SFR, MASR).
- **Abstraction Agreement Score:** The mutual information between clusters produced by different abstraction layers.
- **Disambiguation Speed:** The number of steps required to correctly classify the co-player's policy using the agent's latent representation.
- **Zero-shot Transfer:** Generalization of $\Delta$SEC behavior to novel co-player policies not seen during training.
- **Influence Utility:** The degree to which the agent's actions causally shift the co-player's behavior in predictable (compressible) ways.

We supplement these with qualitative trajectory analysis, attention weight visualizations, and embedding space clustering over time.

### 4.4 Curriculum Over Influence Ambiguity

Inspired by embodied auto-curriculum methods, we implement an influence-level curriculum. Co-players are selected not for novelty or reward challenge, but for the marginal gain they induce in the agent's abstraction refinement. Early training emphasizes maximally ambiguous or misleading co-players (e.g., mimicry-based or counterfactual policies), while later stages increase structural depth (e.g., hidden influence on environment dynamics).

This process ensures that the agent is consistently learning from co-players who are at its current abstraction frontier—supporting open-ended strategic growth rather than overfitting to any specific partner model.

## 5 Addressing Limitations and Future Improvements

### 5.1 Empirical Validation

While the theoretical motivation for Strategic Open-Endedness and ΔSEC is strong, a lack of experimental validation remains a central limitation. To address this:

- **Simplified environments:** We propose beginning with lightweight multi-agent testbeds such as matrix games, simplified gridworlds, or single-skill Overcooked layouts. These environments allow rapid prototyping and fine-grained control over co-player behavior and ambiguity.

- **Progressive layer validation:** Each abstraction layer (IC, CSC, SFR, MASR) will be introduced incrementally. For example, evaluating disambiguation accuracy and behavior prediction in IC before deploying SFR or MASR.

- **Benchmark comparisons:** Agents trained with ΔSEC will be compared against those using standard intrinsic motivation signals (e.g., curiosity, empowerment, RND). These baselines will contextualize the gains offered by influence-grounded learning.

### 5.2 Computational Complexity

The framework's multi-layer structure raises concerns around scalability and computational overhead. Several design interventions can mitigate this:

- **Hierarchical approximations:** We will implement approximate variants of SFR and MASR that activate only when earlier layers (e.g., IC or CSC) detect high uncertainty or strategic novelty.

- **Amortized inference:** Strategic encodings can be produced via learned variational approximations (e.g., amortized VAEs), reducing runtime cost during inference.

- **Sparse updates:** Higher-complexity models will update on a slower temporal schedule, or be conditionally triggered based on thresholded novelty or entropy metrics.

### 5.3 Emergence Guarantees

While open-endedness by nature resists full formalization, safeguards are needed to avoid premature convergence or homogenization of strategy.

- **Theoretical analysis:** We aim to develop convergence and diversity proofs in simplified cases (e.g., discrete Markov games with finite SEC partitions), showing that ΔSEC intrinsically rewards strategic disambiguation.

- **Adversarial meta-agents:** These entities are trained to detect and exploit population-level homogeneity, providing pressure against convergence and encouraging continual abstraction refinement.

- **Diversity bonuses:** In addition to ΔSEC, we include entropy-based diversity regularization across the agent population when SEC distributions collapse.

### 5.4 Human Alignment

Strategic abstraction becomes most valuable when aligned with human interpretability and intent. To close the alignment gap:

- **Human-in-the-loop shaping:** We will integrate sparse human feedback on co-player disambiguation, intention modeling, and abstraction clustering. This feedback will inform preference shaping or reward calibration.
- **Value alignment abstraction layer:** We propose a fifth abstraction layer that learns embeddings optimized for alignment with human-labeled strategic distinctions (e.g., cooperative vs. deceptive).
- **Interpretability tools:** Visualization interfaces will be developed to expose real-time SEC assignments, influence trajectories, and latent model predictions, allowing humans to inspect and shape the learning process.

### 5.5 Additional Research Directions

Beyond resolving limitations, the framework opens new lines of inquiry:

- **Partial observability:** Extending SEC-based abstraction to POMDP settings introduces a new frontier for strategic inference under epistemic uncertainty.
- **Transfer learning:** We will explicitly test the reusability of strategic embeddings across environment configurations, co-player pools, and layouts.
- **Scalability to large-agent settings:** We plan to prototype scalable SEC abstraction in settings with 10–100 agents, where group-level structure and local influence pathways dominate.
- **Theoretical integration:** Future iterations will strengthen ties to game-theoretic equilibrium refinements, multi-agent learning literature, and cognitive models of social inference and joint attention.

These improvements form the next phase of this project, turning the foundational theory into a robust and generalizable toolkit for building strategically competent agents in complex, open-ended systems.

## 6 Ablation Studies

To isolate the contributions of each abstraction layer in the USRLT framework and to test the necessity of influence-driven intrinsic motivation, we design a suite of ablation studies. These experiments selectively remove, substitute, or perturb specific components of the agent's architecture or training signal, allowing us to evaluate the relative importance of each to the emergence of strategic open-endedness.

Our ablation methodology is structured around three key axes:

- **Layer-wise removal:** Eliminate one of the four abstraction layers (IC, CSC, SFR, MASR) and measure the impact on both disambiguation and strategic alignment.
- **Reward perturbation:** Replace the $\Delta$SEC-based intrinsic reward with standard alternatives (e.g., curiosity, empowerment, RND) to test whether general intrinsic motivation is sufficient to drive strategic abstraction.
- **Curriculum interference:** Remove or scramble the influence-level curriculum to assess whether emergent abstraction requires a progression of ambiguous co-players.

## 6.1 Ablation 1: Layer Dropouts

In this set of experiments, we retrain agents with only a subset of the USRLT layers active. Each condition removes one of the four modules (e.g., IC), while keeping the remaining components and training objective intact.

- **Evaluation:** We measure the degradation in $\Delta$SEC, cluster coherence, and alignment across the remaining abstraction layers.
- **Hypothesis:** We expect IC and CSC to be critical for early disambiguation, while SFR and MASR contribute to long-term generalization and strategic forecasting.

## 6.2 Ablation 2: Intrinsic Reward Substitution

We replace the $\Delta$SEC objective with commonly used intrinsic rewards:

- **Curiosity:** Prediction error on forward dynamics.
- **RND:** Error on predicting features from a frozen random network.
- **Empowerment:** Mutual information between actions and future states.

These rewards are computed using the same trajectory buffers and representation backbone to ensure comparability.

- **Evaluation:** Disambiguation speed, influence utility, and generalization to structurally similar but untrained co-players.
- **Hypothesis:** While standard intrinsic rewards may drive exploration, they will not produce coherent strategic compressions or transferable abstractions.

## 6.3 Ablation 3: Curriculum Deactivation

We test two curriculum variants:

1. **Uniform Co-Player Sampling:** No prioritization of ambiguous or frontier co-policies.
2. **Randomized** $\Delta$SEC **Gradient:** Co-player selection is based on noise-corrupted estimates of abstraction gain.

These ablations test the hypothesis that structured exposure to disambiguation pressure is required for efficient abstraction development.

- **Evaluation:** Rate of abstraction emergence over training, and robustness of embeddings under novel co-player dynamics.
- **Expected Result:** Agents without curriculum signal will overfit to a flat SEC landscape, developing brittle or collapsed representations.

## 6.4 Cross-Ablation Summary Table

## 6.5 Interpretation

These ablations clarify that:

- No single layer is sufficient for full strategic abstraction, the framework is synergistic.
- Standard intrinsic rewards may support novelty-seeking but fail to produce compressible or transferable co-agent models.
- Influence-level curricula are essential for agents to encounter and resolve representational ambiguity over time.

| **Ablation Condition** | ΔSEC | Alignment Score | Transferability | Strategic Utility |
|---|---|---|---|---|
| Full USRLT Agent | ✓✓✓ | ✓✓✓ | ✓✓✓ | ✓✓✓ |
| − IC (Intent Certainty) | ✓ | ✗ | ✗ | ✓ |
| − CSC (Contrastive Coding) | ✓ | ✓ | ✗ | ✓ |
| − SFR (Successor Features) | ✓✓ | ✓ | ✓ | ✗ |
| − MASR (Model-Aware) | ✓✓ | ✓✓ | ✓ | ✗ |
| → RND Reward | ✗ | ✗ | ✗ | ✗ |
| → Curiosity Reward | ✓ | ✗ | ✗ | ✓ |
| → Empowerment Reward | ✓ | ✓ | ✗ | ✓ |
| No Curriculum | ✓ | ✗ | ✗ | ✓ |
| Randomized Curriculum | ✓ | ✓ | ✗ | ✓ |

Table 1: Summary of ablation impacts across key performance dimensions.

### 6.6

Our core hypothesis is that agents trained under the **Strategic Open-Endedness** framework will not only develop richer internal representations of co-player behavior, but will also generalize more fluidly across interaction regimes that require implicit coordination, adversarial modeling, or structural inference. We outline our expected results below, grouped by empirical axis and theoretical implication.

## 6.7 Expected Results

**1. Emergent Abstraction via ΔSEC** We expect the agent's representation space to evolve in a structured, layered fashion:

- Early training will emphasize disambiguation of surface-level behavior (Intent Certainty, CSC), while later training will drive convergence in deeper layers (SFR, MASR).
- ΔSEC($t$) will exhibit characteristic deceleration: steep early entropy reduction followed by plateauing as co-player models stabilize.

**2. Cross-Layer Alignment**

- Representations learned by IC, CSC, SFR, and MASR will converge onto a shared latent space, evidenced by high mutual information across layer-specific clusterings.
- This alignment will not occur in baseline agents (e.g., RND, Curiosity, Empowerment), which may form behaviorally salient but strategically brittle partitions.

**3. Zero-Shot Strategic Generalization**

- When exposed to novel co-players (unseen policy families), agents trained with ΔSEC will show high transfer in both compression efficiency and strategic adaptation.
- We anticipate a marked drop in performance from baselines under the same conditions, as their representations are tied to observable novelty rather than latent influence structure.

**4. Curriculum-Driven Growth in Representational Fidelity**

- Influence-level curricula will drive progressive improvements in strategic modeling efficiency — measured as fewer steps to achieve posterior certainty over co-player class.
- Agents trained without a curriculum will stagnate earlier, exhibiting slower disambiguation and poorer alignment across layers.

# 7 Expected Results and Implications

Our core hypothesis is that agents trained under the **Strategic Open-Endedness** framework will not only develop richer internal representations of co-player behavior, but will also generalize more fluidly across interaction regimes that require implicit coordination, adversarial modeling, or structural inference. We outline our expected results below, grouped by empirical axis and theoretical implication.

## 7.1 Expected Results

**1. Emergent Abstraction via ΔSEC** We expect the agent's representation space to evolve in a structured, layered fashion:

- Early training will emphasize disambiguation of surface-level behavior (Intent Certainty, CSC), while later training will drive convergence in deeper layers (SFR, MASR).
- $\Delta\text{SEC}(t)$ will exhibit characteristic deceleration: steep early entropy reduction followed by plateauing as co-player models stabilize.

**2. Cross-Layer Alignment**

- Representations learned by IC, CSC, SFR, and MASR will converge onto a shared latent space, evidenced by high mutual information across layer-specific clusterings.
- This alignment will not occur in baseline agents (e.g., RND, Curiosity, Empowerment), which may form behaviorally salient but strategically brittle partitions.

**3. Zero-Shot Strategic Generalization**

- When exposed to novel co-players (unseen policy families), agents trained with ΔSEC will show high transfer in both compression efficiency and strategic adaptation.
- We anticipate a marked drop in performance from baselines under the same conditions, as their representations are tied to observable novelty rather than latent influence structure.

**4. Curriculum-Driven Growth in Representational Fidelity**

- Influence-level curricula will drive progressive improvements in strategic modeling efficiency — measured as fewer steps to achieve posterior certainty over co-player class.
- Agents trained without a curriculum will stagnate earlier, exhibiting slower disambiguation and poorer alignment across layers.

# 8 Future Work and Closing Remarks

## 8.1 Future Work

While this proposal establishes the foundations for Strategic Open-Endedness, several promising directions remain unexplored. These fall into three categories: representational scaling, agent-agent interaction complexity, and interface with human feedback.

**1. Scaling Abstraction Beyond Pairwise Influence** Our current framework treats strategic abstraction largely as a function of dyadic (ego ↔ co-player) interactions. Future work should extend this to multi-party abstraction, where SECs emerge from coalitional dynamics or distributed influence patterns. This would enable agents to model complex interdependencies, such as emergent group norms or adversarial subgroups.

**2. Latent Structure Discovery in Non-Strategic Environments** Can an agent apply strategic abstraction in environments where no explicit agent is present, but latent structure exists (e.g., market dynamics, ecological feedback, or regulatory constraints)? By framing any causally entangled process as a "virtual co-agent," we may extend this theory to broader classes of problems—including economics, governance, and sustainability.

**3. Influence-Legibility for Human-AI Alignment** An underexplored axis of alignment research is not how much influence an agent has, but how interpretable its influence is to others. Future work can explore how agents might be trained to optimize not just compressibility of influence internally, but communicability of influence externally—creating policies that are strategically expressive to both humans and other agents.

**4. From Influence Recognition to Norm Induction** Beyond modeling the strategies of others, agents may learn to induce norms—persistent strategic attractors—through consistent behavior in shared environments. We aim to explore how SECs may serve not only as a compression tool, but as a scaffold for emergent conventions and proto-institutions among adaptive agents.

**5. Unifying Strategic Open-Endedness with Planning and Memory** We anticipate extending this framework by integrating strategic abstraction with model-based planning and episodic memory. Agents could learn not only what others might do, but how and when to recall specific strategic contexts from their past. This may be key for continual adaptation in temporally extended social ecosystems.

## 8.2 Closing Remarks

This proposal reframes open-endedness not as a challenge of exploration, but of strategic inference. We argue that robustness, generalization, and cooperation emerge not from exhaustive search or extrinsic optimization, but from an agent's ability to compress, refine, and act upon abstract models of influence.

By implementing the Unified Strategic Representation Learning Theory (USRLT) within the MettaGrid environment and grounding learning in the reduction of ambiguity over co-player behavior, we provide both a conceptual and technical foundation for a new class of adaptive agents—not reward maximizers, but structure discoverers.

In doing so, we hope to contribute to a broader shift in how we think about artificial intelligence: away from goal specification, and toward the emergent understanding of goals, agents, and constraints through strategically grounded interaction.

*To learn in open-ended worlds, agents must not only move and observe—they must abstract, influence, and be understood.*

# 9 Annotated Bibliography

This annotated bibliography outlines foundational and recent works that underpin the Strategic Open-Endedness framework and the Unified Strategic Representation Learning Theory (USRLT). The references span key domains: strategic abstraction, causal reasoning, theory of mind, goal-directedness, intrinsic motivation, and safe exploration in multi-agent systems.

## Strategic Abstraction and Representation Learning

### [Tanwisuth 2025] — A Unified Theory of Strategic Representations

This work introduces a hierarchy of strategic abstraction layers—Intent Certainty (IC), Contrastive Strategic Coding (CSC), Successor Feature Strategic Similarity (SFR-SS), and Model-Aware Strategic Similarity (MASR-SS)—to define what an agent minimally needs to understand about others for effective coordination. Strategic Equivalence Classes (SECs) are defined as behaviorally grounded compressions that retain only what influences best responses. A central contribution is formalizing ΔSEC as a learning signal for abstraction refinement.

### [Lauffer et al. 2023] — Who Needs to Know? Minimal Knowledge for Optimal Coordination

This paper defines Strategic Equivalence Relations (SERs) and valued variants (VSERs) that identify the minimal distinctions needed between co-player policies to enable optimal coordination. These equivalence classes are computable in Dec-POMDPs and have practical utility in reducing policy modeling complexity.

### [Oord, Li, and Vinyals 2019] — Contrastive Predictive Coding

Introduces InfoNCE and CPC, an unsupervised representation learning method that preserves predictive structure across time. Forms the backbone of the CSC abstraction layer in USRLT by enabling learning of representations aligned with future behavior prediction.

### [Barreto et al. 2018] — Successor Features for Transfer in Reinforcement Learning

Defines Successor Features (SFs) and Generalized Policy Improvement (GPI), allowing agents to generalize value estimates across reward functions. The SF+GPI framework is foundational for the SFR-SS layer in USRLT.

### [Reinke and Alameda-Pineda 2022] — Successor Feature Representations

Extends SF to Successor Feature Representations (SFRs) by removing the linear reward assumption. This enables value decomposition over non-linear and continuous reward settings, increasing the generality of strategic outcome modeling.

## Causality, Influence, and Theory of Mind

### [Foxabbott et al. 2024] — A Causal Model of Theory-of-Mind in AI Agents

By introducing II-MAIDs (Incomplete-Information Multi-Agent Influence Diagrams), this work extends theory-of-mind modeling to include agents with false or mismatched beliefs. This is critical for analyzing deception, misalignment, and epistemic safety. II-MAIDs unify recursive beliefs and causal reasoning within game-theoretic models.

### [Duéñez-Guzmán et al. 2023] — A Social Path to Human-Like Artificial Intelligence

This perspective argues that intelligence and open-ended learning emerge from multi-scale, multi-agent social interactions. It introduces *compounding innovation*—a synergy of exploration and exploitation that emerges via arms races, population pressures, social relationships, and major transitions. The authors present autocurricula as mechanisms for continual data enrichment and emphasize the role of cooperative and competitive incentives in shaping the learning trajectories of both individual and collective agents. This work supports the claim that strategic representation and influence modeling are foundational to generalizable intelligence, especially when situated within evolving agent collectives.

## Goal-Directedness and Interpretability

### [MacDermott et al. 2024] — Measuring Goal-Directedness

Proposes MEG (Maximum Entropy Goal-directedness), a formal measure of goal-pursuit based on the informativeness of policy behavior relative to hypothetical utility functions. This diagnostic tool complements structural abstraction by quantifying policy intentionality across task regimes.

### [Howard 1974] — General Metagames

A foundational theoretical framework that models recursive mutual prediction through metagame trees. These general metagames form the philosophical and mathematical precursor to SER and II-MAIDs by formally defining strategic foresight and coordination stability.

## Intrinsic Motivation and Reward Shaping

### [Lidayan, Dennis, and Russell 2024] — BAMDP Shaping

Builds a formal reward shaping framework over Bayes-Adaptive MDPs. Introduces BAMDP Potential-Based Shaping Functions (BAMPFs) to unify and correct flawed intrinsic rewards while preserving optimality and safety. A key building block for safe, exploration-driven agents.

### [Jaques et al. 2019] — Social Influence as Intrinsic Motivation

Proposes influence-based intrinsic rewards grounded in causal counterfactuals. Using KL divergence between actual and simulated responses, this work provides a scalable mechanism for emergent coordination, with influence aligned to both legibility and utility in multi-agent environments.

## Foundational Unification and Theoretical Scaffolding

### [Hughes et al. 2024] — Open-Endedness is Essential for Artificial Superhuman Intelligence

This position paper offers a formal observer-relative definition of open-endedness, identifying it as the continual generation of artifacts that are both *novel* and *learnable*. It argues that open-endedness is not incidental but necessary for achieving Artificial Superhuman Intelligence (ASI), and highlights the synergy between foundation models and open-ended algorithms. The notion of "observer surprise with retrospective coherence" operationalizes the challenge of scaling creativity safely. This work motivates grounding open-endedness in interaction-based learning, and underscores the importance of evolving curriculum, social complexity, and meta-generative processes—aligning strongly with the strategic reframing in this proposal.

### [Hammond et al. 2023] — Reasoning about Causality in Games

This paper formalizes causal games and structural causal games (SCGs) by extending MAIDs to support interventions and counterfactuals. Mechanized MAIDs capture how strategies causally influence each other, enabling counterfactual analysis of strategic interactions—useful for safety, fairness, and interpretability.

### [Elias Bareinboim 2024] — An Introduction to Causal Reinforcement Learning

This monograph introduces Causal Decision Models (CDMs) and Causal Reinforcement Learning Tasks, unifying Pearl's causal hierarchy with RL settings. It formalizes how and when observational, interventional, or counterfactual data are needed for learning—critical for epistemically aligned exploration.

**[Levine 2018] — Reinforcement Learning and Control as Probabilistic Inference**

Frames RL as structured variational inference over optimal trajectories. This tutorial undergirds maximum entropy RL, generative skill learning, and soft value functions. It provides a unifying lens for modeling behavior as probabilistic, goal-conditioned inference.

**[Dennis et al. 2020] — Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design**

Introduces the PAIRED (Protagonist Antagonist Induced Regret Environment Design) algorithm, a novel approach to Unsupervised Environment Design (UED). PAIRED uses a triadic game involving a protagonist agent, an antagonist agent, and an adversary that generates environments to maximize regret—defined as the performance gap between the two agents. This drives the emergence of complex, curriculum-shaped tasks that remain solvable, avoiding the failure modes of domain randomization and minimax training. The paper frames PAIRED as an approximation to a minimax regret decision rule and shows it produces agents with higher zero-shot transfer capabilities. It also formalizes UED as a generalization of decision-making under uncertainty, connecting reinforcement learning with classical decision theory. This work is foundational to curriculum generation, emergent complexity, and safe transfer in open-ended agent training.

This annotated bibliography represents a curated set of works that form the theoretical and computational backbone for strategic abstraction, causal modeling, influence reasoning, and intrinsic motivation in open-ended multi-agent environments. Collectively, they scaffold the development of introspective, influence-aware, and strategically generalizable artificial agents.

# References

Barreto, André et al. (Apr. 2018). *Successor Features for Transfer in Reinforcement Learning.* arXiv:1606.05312 [cs]. DOI: `10.48550/arXiv.1606.05312`. URL: `http://arxiv.org/abs/1606.05312` (visited on 02/18/2025).

Dennis, Michael et al. (2020). “Emergent complexity and zero-shot transfer via unsupervised environment design”. In: *Advances in neural information processing systems* 33, pp. 13049–13061.

Duéñez-Guzmán, Edgar A et al. (2023). “A social path to human-like artificial intelligence”. In: *Nature machine intelligence* 5.11, pp. 1181–1188.

Elias Bareinboim Junzhe Zhang, Sanghack Lee (2024). *An Introduction to Causal Reinforcement Learning.* Preprint. URL: `https://causalai.net/r65.pdf`.

Foxabbott, Jack et al. (2024). “A Causal Model of Theory-of-Mind in AI Agents”. In.

Hammond, Lewis et al. (2023). “Reasoning about causality in games”. In: *Artificial Intelligence* 320, p. 103919.

Howard, Nigel (1974). “‘General’metagames: an extension of the metagame concept”. In: *Game Theory as a Theory of a Conflict Resolution.* Springer, pp. 261–283.

Hughes, Edward et al. (June 2024). *Open-Endedness is Essential for Artificial Superhuman Intelligence.* arXiv:2406.04268 [cs]. DOI: `10.48550/arXiv.2406.04268`. URL: `http://arxiv.org/abs/2406.04268` (visited on 02/17/2025).

Jaques, Natasha et al. (June 2019). *Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning.* arXiv:1810.08647 [cs]. DOI: `10.48550/arXiv.1810.08647`. URL: `http://arxiv.org/abs/1810.08647` (visited on 02/26/2025).

Lauffer, Niklas et al. (July 2023). *Who Needs to Know? Minimal Knowledge for Optimal Coordination.* arXiv:2306.09309 [cs]. DOI: `10.48550/arXiv.2306.09309`. URL: `http://arxiv.org/abs/2306.09309` (visited on 02/18/2025).

Levine, Sergey (2018). “Reinforcement learning and control as probabilistic inference: Tutorial and review”. In: *arXiv preprint arXiv:1805.00909.*

Lidayan, Aly, Michael Dennis, and Stuart Russell (2024). “BAMDP shaping: a unified theoretical framework for intrinsic motivation and reward shaping”. In: *arXiv preprint arXiv:2409.05358.*

MacDermott, Matt et al. (2024). "Measuring Goal-Directedness". In: *Advances in Neural Information Processing Systems* 37, pp. 11412–11431.

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (Jan. 2019). *Representation Learning with Contrastive Predictive Coding.* arXiv:1807.03748 [cs]. DOI: `10.48550/arXiv.1807.03748`. URL: `http://arxiv.org/abs/1807.03748` (visited on 02/18/2025).

Reinke, Chris and Xavier Alameda-Pineda (Dec. 2022). "Successor Feature Representations". en. In: *Transactions on Machine Learning Research.* ISSN: 2835-8856. URL: `https://openreview.net/forum?id=MTFf1rDDEI` (visited on 03/09/2025).

Tanwisuth, Sandy (2025). *Unified Strategic Representation Learning Theory (USRLT).* Preprint. URL: `https://www.overleaf.com/read/sfdrqbwvrqwm#62f286`.